第 12 章

資料表正規化

本章提要

- ■1.實體 關係圖與正規化的關係
- 2. 欄位相依
- 3.第一階正規化 (1NF)
- 4. 第二階正規化 (2NF)
- 5. 第三階正規化 (3NF)
- 6.Boyce-Codd 正規化 (BCNF)
- 7. 正規化的另類思考

1. 實體 - 關係圖與正規化的關係

- 資料表的正規化理論 (Normalization Theory) 其實是一套資料表分割的法則
- 在資料庫設計之初,我們可將所有的資料 欄位合併成一個大資料表
- 然後依尋正規化理論所提出的各個步驟, 逐步分割,進而形成許多獨立、但彼此關 聯的小資料表

資料庫的正規化

- ■正規化的目的
 - 要避免資料**重複或相互矛盾**的情形,並使資料車在使用時能更有效率、更容易維護
- ■注意:必須付出增加查詢時間的代價



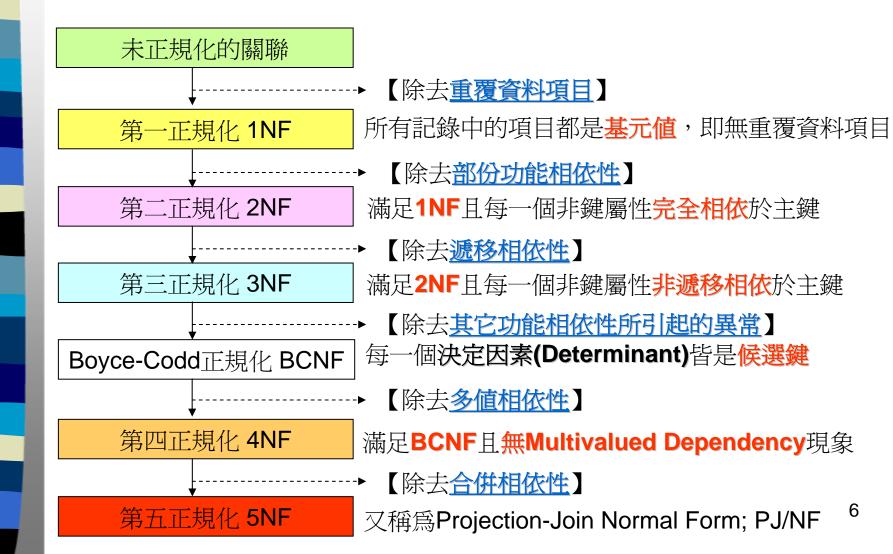
單一的大資料表

多個獨立但相關聯的小資料表

資料庫的正規化

- 資料庫的正規化共可分為 1NF (Normal Form)、2NF、3NF、BCNF (Boyce-Codd Normal Form)、4NF、5NF 等多個階段
- 不過對於一般資料庫設計來說,通常只需要執 行到 BCNF 即可
- 在資料表正規化的過程 (1NF 到 BCNF) 中, 每個階段都是以欄位的相依性, 做為分割資料表的依據之一

正規化步驟



2. 欄位相依

- 在一個資料表中,若乙欄位的值必須搭配甲欄位才有意義,則我們說『乙欄位相依於甲欄位』。
 - 舉例來說, 一個員工資料表如下:

* 員工編號	姓名	地址
101	陳鐵雄	***
102	王大明	****

 對於函數相依X→Y , 我們稱 X決定Y 或Y 函 數相依於X

欄位相依

- 員工編號欄爲主鍵, 做爲唯一辨識該筆記錄的欄位。
 - 姓名欄必須要相依於員工編號欄,對此資料 表來說,姓名欄才有意義
 - 否則同名同姓的陳鐵雄都可對應到該筆記錄
 - 地址欄亦必須相依於員工編號欄, 才有意義, 例如夫妻

欄位相依

■ 我們再看看下面的例子:

*課程編號	*學號	成績
M102	903123	89
A205	903145	75
	*	
	欄位相依	

欄位相依

- ■成績欄本身如果單獨存在是沒有意義的, 因為不曉得是哪一門課、哪個學生的成績
 - 當成績欄相依於課程編號及學號二欄時,就可以了解某個學生修某堂課的成績,這樣的成績資料才有意義
 - 課程編號及學號二欄決定成績欄

3. 第一階正規化 (1NF)

- ■正規化的過程是循序漸進的
 - 資料表必須在滿足第一階正規化的條件之下, 才能進行第二階正規化
 - 也就是說,第二階正規化必須建立在符合第一階正規化的資料表上,依此類推
- ■第一階正規化正是所有正規化的基礎

3. 第一階正規化 (1NF)

- ■第一階正規化的規則
- 未符合 1NF 資料表的缺點
- 建構 1NF 資料表的方法

第一階正規化的規則

- 如果資料表符合以下條件,我們就說這個資料表符合第一階正規化的形式 (First Normal Form, 簡稱 1NF):
 - 1. 資料表中有主鍵, 而其他所有的欄位都相依於主鍵。
 - 2. 每個欄位中都只有儲存**單一值** 例如,姓名欄位中不能存放 2 個人的姓名。
 - 3. 資料表中沒有**意義相同的多個欄位** 例如,姓名1、姓名2...等重複的欄位。

■ 首先我們來看一個非正規化的資料表:

選課資料表

課程編號	課程名稱	教師編號	教師姓名	學場	學生姓名	成績
CS101	資料庫系統	1001	曾聰明	91010 91013 91020	陳鐵雄 王大明 李珍珍	70 90 88
CS103	影像處理	1015	郝天才	91033 91045	進阿丁	86 93

- 很明顯的,此資料表違反了第一階正規化的第 2 個條件
- ■可能有兩個老師同時都開了『資料庫系統』的課程,也可能有兩個學生都叫做 『王大明』
- 這個資料表缺少具有唯一性的主鍵, 也違 反了第一階正規化的第 1 個條件

1. 『學號』、『學生姓名』及『成績』欄 的長度無法確定 由於修課學生的人數可多可少,所以必 須預留很大的空間給這兩個欄位,如此 反而造成儲存空間的浪費

2. 降低存取資料的效率

例如要找出『陳鐵雄』的成績,必須先在學生姓名欄中找出『陳鐵雄』所在的位置,然後才能從成績欄中擷取出對應的成績資料,這不僅減緩了資料處理的速度,而且也增加了程式出錯的機會

■接下來看一個違反第一階正規化第 3 個條件的例子:

修課資料表

課程名稱	學生 1	學生 2	學生 3	戏績 1	成績2	成績3
資料庫系統	陳鐵雄	王大明	李珍珍	70	90	88
影像處理	趙阿丁	奠邓可龍		86	93	

- ■像學生1、學生2、學生3 這樣一群意義相同的欄位,其問題同樣是無法確定要有 多少個重複的欄位,而且存取效率低落
- ■例如要找『陳鐵雄』的成績,必須在學生 群組的每一個欄位中搜尋,找到後還得要 到成績群組中的相同位置欄位中讀取,相 當麻煩

建構 1NF 資料表的方法

2000年1000年

■對於不具第一階正規化形式的資料表,我們可將**重複的資料項**分別儲存到不同的記錄中,並加上適當的主鍵:

ě	进床自科女						
	*課程編號	*學號	課程名稱	教師編號	教師姓名	學生姓名	成績
	CS101	91010	資料庫系統	1001	曾聰明	陳鐵雄	70
	CS101	91013	資料庫系統	1001	曾聰明	王大明	90
	CS101	91020	資料庫系統	1001	曾聰明	李珍珍	88
	CS103	91033	影像處理	1015	郝天才	趙阿丁	86
	CS103	91045	影像處理	1015	郝天才	奠邓可龍	93

建構 1NF 資料表的方法

- ■如此一來,雖然增加了許多記錄,但每一個欄位的長度及數目都可以固定
- 我們可用課程編號欄加上學號欄做為主 鍵, 那麼在查詢某學生修某堂課的成績時, 就非常方便而快速了

4. 第二階正規化 (2NF)

- 在執行符合第一階正規化的資料表時,應 該會發覺:我們輸入了許多重複的資料
 - 不但浪費儲存的空間, 更容易造成新增、删除或更新資料時的異常狀況
- 我們必須進行**第二階正規化**,來消除這些問題。

第二階正規化 (2NF)

- ■第二階正規化的規則
- 建構 2NF 資料表的方法

第二階正規化的規則

- 如果資料表符合以下的條件,我們說這個資料表符合第二階正規化的形式 (Second Normal Form,簡稱 2NF):
 - 1. 符合 1NF 的格式。
 - 2. 各欄位與主鍵間沒有部分相依的關係。

第二階正規化的規則

- 部分相依只有在主鍵是由多個欄位組成時才會發生,它是指某些欄位只與主鍵中的部分欄位有相依性,而與另一部分的欄位沒有相依性
- ■以<mark>選課</mark>資料表來說,其主鍵爲**課程編號+** 學號欄位
 - 課程名稱欄只和課程編號欄有相依性
 - 學生姓名欄只和學號欄有相依性

第二階正規化的規則

只和學號有相依性

選課資料表

*課程編號	* 學號	課程名稱	教師編號	教師姓名	學生姓名	成績
CS101	91010	資料庫系統	1001	曾聰明	陳鐵雄	70
CS101	91013	資料庫系統	1001	曾聰明	王大明	90
CS101	91020	資料庫系統	1001	曾聰明	李珍珍	88
CS103	91033	影像處理	1015	郝天才	趙阿丁	86
CS103	91045	影像處理	1015	郝天才	鄭阿甫	93

只和課程編號有相依性

部分相依的問題(1/4)

■新增資料時:

- 若有一個新來的轉學生『吳技安』,但還沒有選修任何課程,那麼它的資料將無法輸入 (因爲主鍵中的欄位值是不允許有空白的,但 此時根本沒有課程編號可輸入)

部分相依的問題(2/4)

■ 更改資料時:

- 當我們想要將課程名稱『資料庫系統』更改 爲『資料庫管理』時,必須搜尋整個資料庫 並一一更改,非常沒有效率

部分相依的問題(3/4)

■刪除資料時:

- 由於陳鐵雄只修了『CS101』一門課,如果 將該筆記錄刪除,那麼陳鐵雄的資料也就跟 著消失了

部分相依的問題(4/4)

- ■部分相依也會造成資料重複出現的問題
 - 例如『CS101, 資料庫系統』這組資料每次都必須同時輸入, 不但浪費時間及儲存空間, 而且也容易因疏忽而造成資料不一致的錯誤。例如:

CS101	資料庫系統	★ 輸入錯誤造成資料不一致
CS101	資料庫管理	-

建構 2NF 資料表的方法

- 要除去**資料表中的部分相依性**, 只需將部份相依的欄位分割成另外的資料表即可
- 例如我們將選課資料表分割成 3 個較小的資料表 (加 "*" 號的欄位為主鍵):

課程資料表

* 課程編號	課程名稱	教師編號	教師姓名
CS101	資料庫系統	1001	曾聰明
CS121	影像處理	1015	郝天才

建構 2NF 資料表的方法

成績資料表

*課程編號	*學號	成績
CS101	91010	70
CS101	91013	90
CS101	91020	88
CS121	91020	86
CS121	91033	93

學生資料表

*學號	學生姓名
91010	陳鐵雄
91013	工大明
91020	李珍珍
91033	趙阿丁

課程資料表

*課程編號	課程名稱	教師編號	教師姓名
CS101	資料庫系統	1001	曾聰明
CS121	影像處理	1015	郝天才

- ■經過了第二階正規化後的資料表,其實還 存在一些問題:
 - 在課程資料表中,如果新來了一位教師『甄 蒡』,在尚未安排他教授的課程之前,我們無 法輸入該教師的資料
 - 若要刪除影像處理的課程, 勢必會將教授該 門課程的教師一倂刪除
 - 若要更改影像處理課程的名稱,則必須同時 更改多筆記錄,造成不便

第三階正規化 (3NF)

- ■第三階正規化的規則
- 建構 3NF 資料表的方法

第三階正規化的規則

- 如果資料表符合以下條件, 我們就說這個 資料表符合第三階正規化的形式 (Third Normal Form, 簡稱 3NF):
 - 符合 2NF 的格式。
 - 各欄位與主鍵間沒有間接相依的關係。

第三階正規化的規則

- ■間接相依是指在二個欄位間並非直接相 依, 而是借助第三個欄位來達成資料相依 的關係
 - 例如 A 相依於 B; 而 B 又相依於 C, 如此 A與 C 之間就是間接相依的關係
- 要找出各欄位與主鍵間的間接相依性, 最 簡單的方式就是看看資料表中有沒有

『與主鍵無關的相依性』存在

第三階正規化的規則

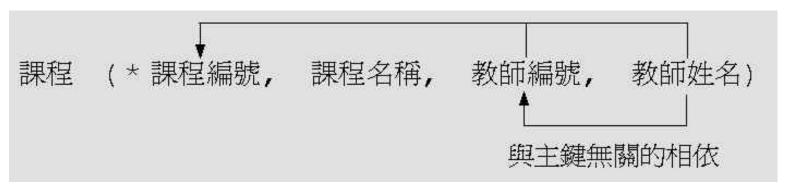
■例如在課程資料表中:

課程資料表

*課程編號	課程名稱	教師編號	教師姓名
CS101	資料庫系統	1001	曾聰明
CS121	影像處理	1015	郝天才

第三階正規化的規則

- ■由於每一門課程都會有授課的教師,所以 教師編號欄和教師姓名欄都相依於課程 編號欄
- 教師姓名又同時相依於教師編號欄, 而這個相依性是與主鍵完全無關的:



建構 3NF 資料表的方法

- 要除去資料表中的間接相依性, 其方法和 除去部分相依性完全相同
 - 例如課程資料表可再分割成兩個資料表:

課程資料表

*課程編號	課程名稱	教師編號
CS101	資料庫系統	1001
CS103	影像處理	1015

教師資料表

*教師編號	教師姓名	
1001	曾聰明	
1015	郝天才	

與直覺式的分割技巧做比較

■ 當您設計資料庫一段時間,累積了經驗及 技術後,您便可依照己身的經驗,以直覺 的方式對資料表執行最佳化

正規化	功能相同的直覺式分割法		
1NF: 有主鍵 欄位中只有一個單一值 沒有意義相同的重複欄位	無		
2NF: 除去『部分相依性』	分割『欄位値一再重複』的欄位		
3NF: 除去『間接相依性』	分割『與主鍵無關』的欄位		

6. Boyce-Codd 正規化 (BCNF)

- ■對於大部分資料庫來說,通常只需要執行 到第三階段的正規化即足夠了
- 如果資料表的主鍵是由多個欄位組成的, 則必須再執行 Boyce-Codd 正規化

Boyce-Codd 正規化 (BCNF)

- Boyce-Codd 正規化的規則
- ■檢驗『成績』資料表是否滿足 BCNF 規 範

Boyce-Codd 正規化的規則

- 如果資料表的主鍵只由單一欄位組成,則符合第三階正規化的資料表,亦符合Boyce-Codd 正規化
- 但若資料表的主鍵由多個欄位組成,則資料表只要符合以下條件,我們就說這個資料表符合Boyce-Codd 正規化的形式(Boyce-Codd Normal Form,簡稱BCNF):

Boyce-Codd 正規化的規則

- 符合 2NF 的格式。
- ■各欄位與主鍵沒有間接相依的關係
- 主鍵中的各欄位不可以相依於其他非主 鍵的欄位

檢驗『成績』資料表是否滿足 BCNF 規範

■ 我們利用 Boyce-Codd 正規化的條件,來 檢驗主鍵由多個欄位組成的成績資表:

成績資料表

*課程編號	*學號	成績
CS101	91010	70
CS101	91013	90
CS101	91020	88 }
CS121	91020	86
CS121	91033	93

所有欄位都相依於主鍵, 主鍵中的欄位亦無相依 於其他非主鍵欄位

檢驗『成績』資料表是否滿足 BCNF 規範

- ■成績資料表是符合『Boyce-Codd 正規化的形式』的資料表
 - 成績欄相依於課程編號及學號欄
 - 對課程編號欄而言, 並無相依於成績欄
 - 對學號欄而言, 也無相依於成績欄

檢驗『成績』資料表是否滿足 BCNF 規範

課程資料表

* 課程編號	課程名稱	教師編號
CS101	資料庫系統	1001
CS103	影像處理	1015

教師資料表

*教師編號	教師姓名
1001	曾聰明
1015	郝天才

成績資料表

*課程編號	*學號	成績
CS101	91010	70
CS101	91013	90
CS101	91020	88
CS121	91020	86
CS121	91033	93

學生資料表

*學號	學生姓名
91010	陳鐵雄
91013	王大明
91020	李珍珍
91033	趙阿丁

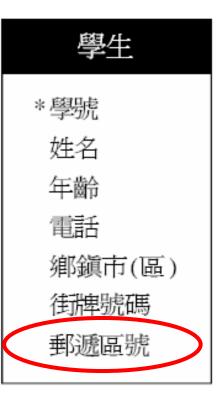
7. 正規化的另類思考

- ■不必要的分割
- ■人工的分割

7. 正規化的另類思考

■不必要的分割

- 正規化的工作有時不必做的非常徹底,因爲過多的資料表可能會降低系統執行的效能
- 例如,以學生資料表來看:



正規化的另類思考

■按照正規化的原理,郵遞區號其實是相依 於『鄉鎮市(區)』,因此這個資料表可再 進行分割,才能符合正規化的要求。如圖

所示:

學生

*學號

姓名

年齡

電話

鄉鎮市(區)編號

街牌號碼

郵遞區號對照

*鄉鎮市(區)編號 鄉鎮市(區)

郵遞區號

正規化的另類思考

- ■實際作業上,一定是使用整筆的地址資料, 如果將它們分割了,那麼每次查詢地址時, 都要透過資料表的關聯,從郵遞區號對照 資料表中取得郵遞區號欄位
- 當資料量增加到一定程度時,就會明顯感 受到系統效能降低的事實,實非明智之舉

正規化的另類思考

■人工的分割

有時爲了增加資料處理的效率,我們可將已經符合 BCNF 的資料表再做分割

- 例如,一個資料表擁有非常多的欄位,而其中又有許 多欄位根本很少用到
- 就可將這些很少用到的欄位分離出來, 存放到另外一個資料表中

練習一

工時薪資資料表

旗旗公司進銷存系統設計

旗旗公司進銷存系統設計

■請將以下資料表依照 1NF→2NF→3NF 的步驟做正規化 (主鍵請在前面加 "*" 號 表示):

專案名稱	員工姓名	工時	薪資等級	時薪
李氏公司人事系統設計案	賴妙麗	100	С	\$130
李氏公司人事系統設計案	王榮恩	130	Α	\$210
李氏公司人事系統設計案	杜馬份	110	В	\$160

160

90

В

C

\$160

\$130

章海格

趙西泊

練習一解答

■ 1 F

工 事 專 書 員 員 工 工 薪 資 號 稱 號 名 編 姓 等

練習一解答

■ 2F

 事案

 *專案編號

 專案名稱

 工 時 薪 資

 *專 案 編 號

 *員 工 編 號

 工 時

 *員工

 *員工姓名

 工資等級

 時薪

練習一解答

■ 3F

 事 案

 *專 案 編 號

 專 案 名 稱

 工時薪資

 *專案編號

 *員工編號

 工時

 計
 計

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *
 1

 *</

 工資等級編號

 *工資等級

 財薪

練習二

某錄影帶店提供了一張有關錄影帶相關資料的表單如下:

代碼	片名	導演	導演拍片量	主角代碼	主角姓名	扮演角色	片長
(ID)	(Title)	(Director)	(D-QNT)	(Actor-ID)	(Actor-N)	(Role)	(Time)
V0001	我就是我	劉華	12	A0105 A0206	薛紅 張軍	鄭萱庭 洪青文	126
A0002	永恆的一日	陳良	8	A0101 A0101 A0206	王藍 王藍 張軍	張億雯 李敏娟 許遠忌	130
V0003	為什麼	劉華 陳良	12 8	A0105 A0101	薛紅 王藍	劉顯耀 唐文娟	112

- (一)請說明正規化(Normalization)之目的?(5分)
- (二)若上表為一關聯表,它是否符合第一階正規化(INF)?為什麼?(5分)
- (三)請將上表轉化成符合第三階正規化 (3NF) 的關聯表,並註明主鍵 (Primary Keys),若需任何假設,請列出。 (10分)
- (四)請繪出相對之實體關係圖(E-R Diagrams),若需任何假設,請列出。(10分)

練習二解答(1/4)

- (一)去除多餘資訊(data redundancy)與避免資料庫的異常現象(插入、刪除、更新異常)。
- (二)否,紀錄項目並非皆基元值(Atomic value)。如: ID=V0001的Actor-ID、Actor-N、Role皆含有多個值。
- (三)假設每部影片只有一位導演,且每位主角在一部影片中止扮演一個角色。

則上述關聯:

ID→Title, Director, Time

Director → D-QNT

Actor-ID→Actor-N

ID ,Actor-ID→Role

1NF:

<u>代碼</u>	片名	導演	導演拍片量	主角代碼	主角姓名	扮演角色	片長
<u>(ID)</u>	(Title)	(Director)	(D-QNT)	(Actor-ID)	(Actor-N)	(Role)	(Time)
V0001	我就是我	劉華	12	A0105	薛紅	鄭萱庭	126
V0001	我就是我	劉華	12	A0206	張軍	洪青文	126
A0002	永恆的一日	陳良	8	A0101	王藍	張億雯	130
A0002	永恆的一日	陳良	8	A0101	王藍	李敏娟	130
A0002	永恆的一日	陳良	8	A0206	張軍	許遠忌	130
V0003	為什麼	劉良	12	A0105	薛紅	劉顯耀	112
V0003	為什麼	劉良	12	A0101	王藍	唐文娟	112

練習二解答(2/4)

2NF :

影片角色

主角代碼	扮演角色
(Actor-ID)	(Role)
A0105	鄭萱庭
A0206	洪青文
A0101	張億雯
A0101	李敏娟
A0206	許遠忌
A0105	劉顯耀
A0101	唐文娟
	A0105 A0206 A0101 A0206 A0101 A0206 A0105

<u>主角代碼</u> (Actor-ID)	主角姓名 (Actor-N)
A0105	薛紅
A0206	張軍
A0101	王藍
A0105	薛紅
	主監 薛紅

影片

<u>代碼</u>	片名	導演	導演拍片量	片長
<u>(ID)</u>	(Title)	(Director)	(D-QNT)	(Time)
V0001	我就是我	劉華	12	126
A0002	永恆的一日	陳良	8	130
V0003	為什麼	劉良	12	112

主角

練習二解答(3/4)

3NF:

影片角色

<u>代碼</u> (D)	<u>主角代碼</u> (Actor-ID)	扮演角色 (Role)
V0001	A0105	鄭萱庭
V0001	A0206	洪青文
A0002	A0101	張億雯
A0002	A0101	李敏娟
A0002	A0206	許遠忌
V0003	A0105	劉顯耀
V0003	A0101	唐文娟

ID為外鍵,參考影片的ID

Actor-ID為外鍵,參考主角的Actor-ID

練習二解答(4/4)

影片

代碼	片名	導演	片長
<u>(ID)</u>	(Title)	(Director)	(Time)
V0001	我就是我	劉華	126
A0002	永恆的一日	陳良	130
V0003	為什麼	劉良	112

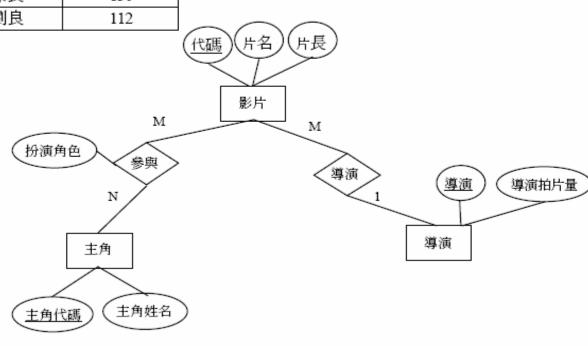
導演為外鍵,參考導演的導演

導演

<u>導演</u>	導演拍片量
(Director)	(D-QNT)
劉華	12
陳良	8
劉良	12

主角

主角代碼	主角姓名
(Actor-ID)	(Actor-N)
A0105	薛紅
A0206	張軍
A0101	王藍
A0105	薛紅



練習三:圖書管理系統

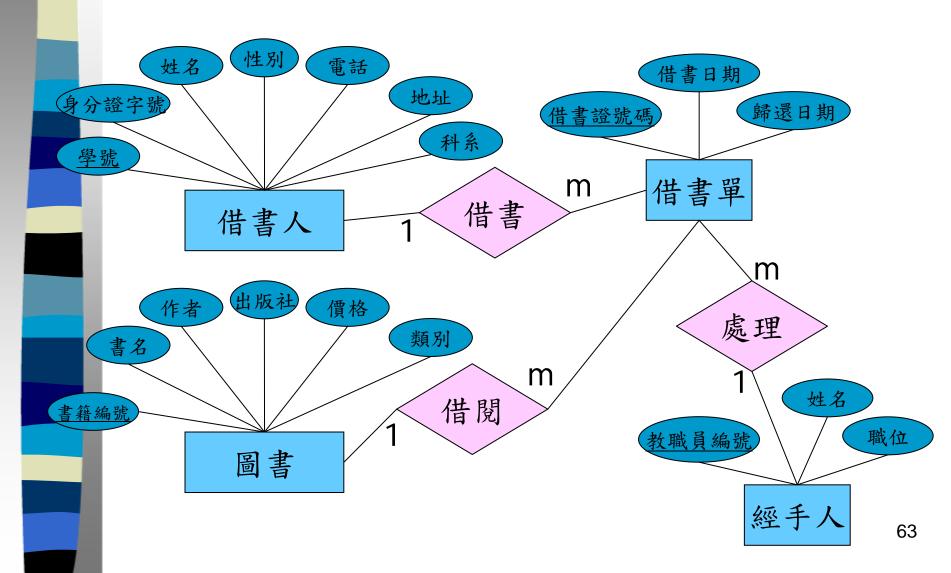
■實體與屬性

- 一借書人:學號、身分證字號、姓名、性別、電話、地址、科系
- 圖書:書籍編號、書名、作者、出版社、價格、類別
- 經手人:教職員編號、姓名、職位
- 一借書單:借書證號碼、借書人、書籍名稱、借書日期、歸還 日期、經手人

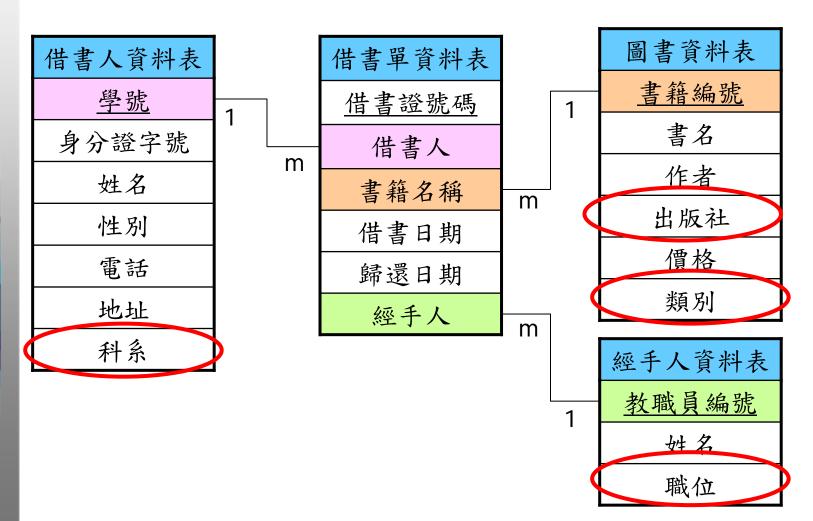
| 關聯

- 「借書人」與「借書單」間為1對多的『借書』關係
- 「經手人」與「借書單」間為1對多的『處理』關係
- 「圖書」與「借書單」間為1對多的『借閱』關係

關聯說明:ERD



關聯說明



關聯說明:ACCESS的第三階正規化

